A Best View Selection in Meetings through Attention Analysis Using a Multi-camera Network

Sebastian Gruenwedel, Xingzhe Xie and Wilfried Philips
Ghent University TELIN-IPI-IBBT
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
{sebastian.gruenwedel, xingzhe.xie}@telin.ugent.be

Chih-Wei Chen and Hamid Aghajan
Stanford University
350 Serra Mall, Stanford, CA, USA
{louistw, aghajan}@stanford.edu

***Abstract*—Human activity analysis is an essential task in ambient intelligence and computer vision. The main focus lies in the automatic analysis of ongoing activities from a multi-camera network. One possible application is meeting analysis which explores the dynamics in meetings using low-level data and inferring high-level activities. However, the detection of such activities is still very challenging due to the often corrupted or imprecise low-level data. In this paper, we present an approach to understand the dynamics in meetings using a multi-camera network, consisting of fixed ambient and portable close-up cameras. As a particular application we are aiming to find the most informative video stream, for example as a representative view for a remote participant. Our contribution is threefold: at first, we estimate the extrinsic parameters of the portable close-up cameras based on head positions. Secondly, we find common overlapping areas based on the consensus of people's orientation. And thirdly, the most informative view for a remote participant is estimated using common overlapping areas. We evaluated our proposed approach and compared it to a motion estimation method. Experimental results show that we can reach an accuracy of 74% compared to manually selected views.**

***Index Terms*—human activity recognition, smart distributed cameras, activity analysis, meeting analysis**

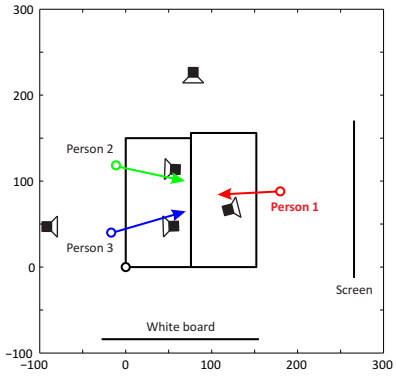
I. Introduction

Human activity analysis is an important area of ambient intelligence and computer vision. The goal of human activity analysis is to automatically analyze ongoing activities from one or multiple unknown video streams. The objective is to correctly classify the video streams into a set of activities [1]. Many applications are available to support people in carrying out their everyday life activities and tasks, such as automatic light control, meeting analysis, etc.

For the latter, making use of low-level data, such as positional data for each meeting attendant [2]–[6] or detailed face analysis [7] could help high-level analysis to understand the dynamics in meetings, for instance. Activities can range from events, like "who is talking" or "who is looking at whom" to more complex ones, such as "who is the main speaker", "who is paying attention in the meeting, who does not", ... However, the detection of such activities is still very challenging. Low-level data is often corrupted or imprecise due to environmental changes. Therefore the low-level data cannot be assumed to be correct or very precise.

In this paper we present an approach to understand the dynamics in meetings in a multi-camera setup, consisting of fixed ambient and portable close-up cameras. We focus on the detection of activities in a certain time period rather than on a frame-by-frame basis. Meeting analysis is a challenging task due to the complexity of detected activities. There are lots of possible applications for meeting analysis, for instance creating a complete protocol of a meeting. In our approach, we focus particularly as an application on detecting the best view of the multi-camera setup to stream this representative view to a remote participant (Fig. 1). The best view hereby refers to as the most informative video stream within the multi-camera setup. This best view is detected by analyzing the head orientation of the participants in the meeting. Our contribution of this approach is threefold: At first, we estimate the extrinsic

(a)



(b)

Fig. 1. *Example*. In Fig. 1a a collection of available video streams are shown in which the most informative stream is chosen (red box). The selection is based on the analysis of participant's orientation in portable close-up cameras (Fig. 1b) for a certain time period. The relation of the participants and their orientation is estimated using fixed ambient cameras.

parameters of the close-up cameras using head positions in the ambient cameras and the corresponding close-up cameras. Our algorithm finds the best set of corresponding points to estimate the extrinsic parameters of a close-up camera. In the second step, the head poses (position and orientation), estimated on a frame-by-frame basis, are used to find common viewing areas where people look at for a certain time period. In the third step these areas are analyzed and used to detect overlapping areas based on the consensus of people in the meeting. Therefore, we can select the most informative view in which an activity takes place.

The paper is structured as follows: In Section II, we discuss related work. Section III describes the proposed approach which consists of three parts: the estimation of the extrinsic parameters of the close-up cameras in Section III-A, finding overlapping area based on the orientation of people in Section III-B and the detection the most informative view for a remote participant in Section III-C. Section IV presents experiments to demonstrate our approach. The results show that we are able to detect the most important view for a remote participant. Section V concludes the paper.

II. Related Work

Determining the rigid motion relating a pair of cameras is a well-studied problem [8] and has been extended to a multiple-camera scenario [9]. Solving this classical problem usually requires a set of matched image correspondences in each of the views. In this work, we consider face locations of occupants in the environment to construct the correspondences because they can be uniquely identified and robustly tracked [10].

Smart meeting systems have been designed [11] to automatically archive, analyze and summarize meetings, which are arguably the most important means of information distribution and exchange. While many proposed smart meeting systems [12], [13] consider sensors of multiple modalities, including audio and video, we focus on visual sensors only, and build a high-level semantic attention detector using features extracted from them.

Head orientation is a good indicator for focus of attention, and can be used to infer social attention and human interaction [14]. In a meeting application scenario, the participants' focus of attention can also be used as an index in an archive [15]. In [15], a panoramic camera captures low resolution images of participants in a round-table meeting, and a Hidden Markov Model (HMM) is employed to estimate the head poses. However, only pan and tilt angles are estimated, and people are assumed to be seated around the table. In [16], a head-mounted eye and head tracking system was used to track the head orientation and gaze direction of a meeting participant. It was shown that head orientation contributes to 68.9% of overall gaze direction, and head orientation alone achieves high accuracy in meeting analysis. In our work, we use a vision-based head tracker to estimate the full six degree-of-freedom head poses with portable cameras. No intrusive sensors are needed, and by recovering the relative position between the portable close-up cameras and the fixed ambient cameras in the meeting room, focus of attention is not restricted to meeting participants only but can be on interesting regions within the environment.

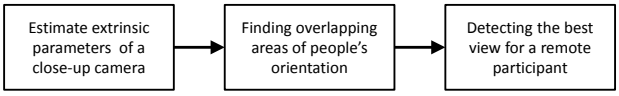


Fig. 2. *Approach overview*. In the first step the extrinsic parameters for the portable close-up cameras are estimated using the fixed ambient cameras and facial information of each close-up camera. In the second step, the head position and orientation are used to find common viewing areas where people look at for a certain time period. Finally, these areas are analyzed and used to detect the most important view in the meeting.

III. Approach Overview

In this section, we describe the proposed approach to understand the dynamics in meetings; in particular to detect the most informative video stream for a remote participant based on a multi-camera setup. The multi-camera setup consists of fixed ambient and portable close-up cameras. There are more possible applications for our proposed approach such as to focus PTZ cameras within a meeting on specific people even for more detailed analysis, or, to automatically report what happened during a meeting.

Figure 2 depicts a block diagram of our proposed approach. First, we estimate the extrinsic parameters for portable close-up cameras. This is needed since close-up cameras are usually laptop-cameras and are placed by the participants themselves. This makes a pre-calculated calibration impossible. Furthermore, participants adjust their cameras or move their cameras during the meeting. Therefore, the calibration of close-up cameras has to be done automatically and instantaneously. In the second step, we find overlapping areas based on the orientation of people, i.e. an overlapping area is defined as an area at which most of the participants are looking. This consensus decision refers to the fact that if people are constantly looking at a certain area, an important activity is happening there. Otherwise, people would not look at this area. In the final step the most informative view for a remote participant is chosen based on the detected overlapping areas of the participants.

A. Estimation of the Extrinsic Parameters of Close-up Cameras

In this section we outline the estimation of the extrinsic parameters for close-up cameras. The extrinsic parameters of a close-up camera are needed to relate the head position and orientation of a participant to a common coordinate system resulting in common viewing areas.

At first, participants are tracked using the calibrated ambient cameras [2], [10]. Making use of multi-camera tracking algorithms enables the estimation of head position $\left\{h_w^i\right\}; i = 1, \ldots, N$ for each participant w.r.t. a common coordinate system w [17]. Assuming we have only one participant per close-up camera, we are tracking the face of the participant in real-time [18] resulting in head positions $\left\{h_c^i\right\}$ w.r.t. the

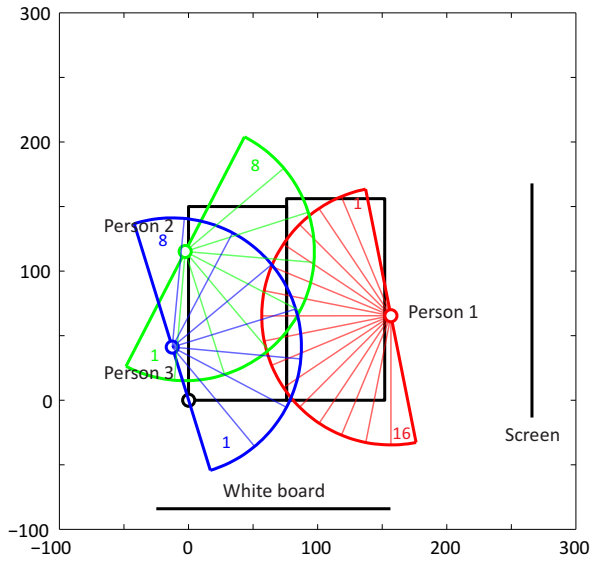


Fig. 3. *Viewing zones.* The semicircles represents N viewing zones for each participant established from the extrinsic parameters of the portable close-up cameras. The head of each participant is the center of each semicircles.

close-up camera coordinate system c. The final task is to find a transformation according to the following equation [19], [20]:

$$h_c^i = R \cdot h_w^i + T + n_i, \tag{1}$$

where R is a rotation matrix, T a translation vector, and n_i a noise vector. Arun *et al.* [19] describe an algorithm to find the least-square solution of R and T based on the Singular Value Decomposition (SVD). Here, the least-square problem is defined as follows:

$$\Sigma^2 = \sum_{i=1}^{N} \| h_c^i - \left(R \cdot h_w^i + T \right) \|^2. \tag{2}$$

In the paper the authors show that, by decoupling the translation from rotation component, it is possible to estimate the least-square solution of R. To do so, the SVD of a matrix H needs to be calculated, where H is defined as follows:

$$H = \sum_{i=1}^{N} \left(h_w^i - \mu_w \right) \left(h_c^i - \mu_c \right)^T, \tag{3}$$

where μ_w is the mean of the point set $\left\{ h_w^i \right\}$ and μ_c the mean of $\left\{ h_c^i \right\}$. But there is a problem with the algorithm. If the noise is too large, it is not possible to find a valid solution for the SVD of H since both sets are coplanar. In this case, a RANSAC based method [21] needs to be used, resulting in a subset of $\left\{ h_w^i \right\}$ and $\left\{ h_c^i \right\}$ to combat against outliers.

B. Finding Overlapping Areas Based on the Orientation of People

By performing head pose estimation [18] and using the extrinsic parameters for a close-up camera, we are able to relate the head position and orientation to a common coordinate system. Therefore, it is possible to find common viewing areas based on the head position and orientation of each participant.

TABLE I
RELATION BETWEEN VIEWING ZONES FOR EACH PERSON AND THE DETECTED ACTIVITIES

Interesting area	Person 1		Person 2		Person 3	
	yaw	pitch	yaw	pitch	yaw	pitch
Person 1	-	-	5	2	4	2
Person 2	5-7	2	-	-	6,7	2
Person 3	8-10	2	2,3	2	-	-
Screen	u	u	2,3	1	1,2	1
White board	11-15	1	4,5	1	3,4	1

The head pose consists of a translation and rotational component w.r.t. the camera coordinate system. The three degrees of freedom of a human head can be described by the egocentric rotation angles *pitch*, *roll*, and *yaw*. Pitch is expressed as turning one's head up or down, yaw means that one person turning his or her head left or right, and roll describes the activity of people moving their heads towards their shoulders. In our approach, we are interested in the head movement and the rotation angles, pitch and yaw, to analyze people's looking behaviors in a meeting.

The head's yaw rotation angle ranges from $[-90°, 90°]$. This results in a viewing area of $180°$ for each participant. We divide this viewing area into N evenly-spaced angular zones. In Figure 3 semicircles represent these N viewing zones for each participant, labeled in a counter-clockwise direction. Note that these semicircles are only used for visualization and the viewing areas are not limited to these semicircles. For example the position of Person 1's head is the center of the semicircle for this person. Within the semicircle the different viewing zones are shown. This semicircle is obtained by the head pose estimation and the calibrated portable close-up camera, described in Section III-A.

Given the yaw rotation angle α, the order of viewing zones z can be calculated using the following formula:

$$z = \text{floor} \left(\frac{\alpha + 90°}{180°} N \right) + 1 \tag{4}$$

The function floor (A) rounds the variable A to the nearest integers less than or equal to A.

Besides, the pitch rotation angle is classified into two zones; *"looking up"* (zone 1) and *"otherwise"* (zone 2).

To understand the dynamics in a meeting we need to detect activities which are important for an application. In our particular application, we are interested in the most informative view for a remote participant. During a meeting, participants usually perform one of the following activities: one person talks and others look at him or her; one person introduces something on the screen or the white board and others look either at the screen or the white board, i.e. there are some specific areas in the meeting room which are of interest for a remote participant. Therefore, as an activity we are looking for the overlapping area of the participants which is defined by the consensus of the people's orientation. Since we have three participants in our meetings, a white board and a screen, we only focus on these five interesting areas in our paper.

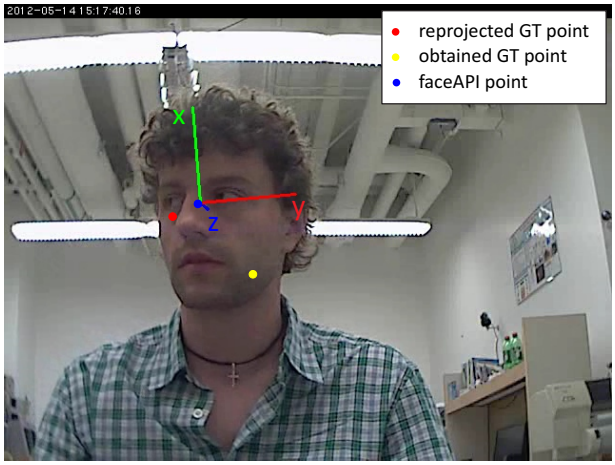


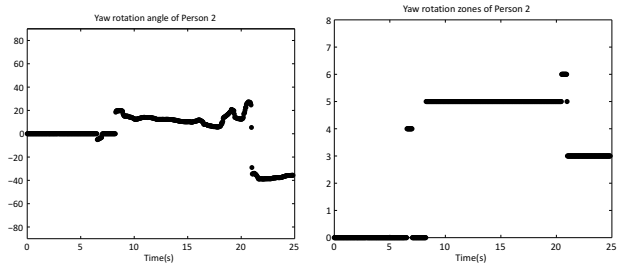
Fig. 4. *Point correspondence for one frame*. In this example, the annotated GT point and the head pose obtained by faceAPI [18] (position and orientation) are used to estimate the extrinsic parameters for a portable close-up camera. Note the noise for this point correspondence is quite large. Therefore, we favor a RANSAC like approach to cope with noisy correspondences.

In Table I the relation between viewing zones of each participant and the interesting areas are described. We use a decision based system to detect an activity. In the table the symbol u means that the viewing zone is unknown. This is due to the fact that Person 1 needs to turn around to see the screen and therefore will not be seen by the close-up camera. The viewing areas of Person 2 and 3 are divided into 8 zones separately. However, due to the proximity of Person 2 and 3 it is challenging to differentiate the looking behavior of Person 1 with 8 zones. Therefore, we divide the viewing area of Person 1 into 16 zones. For Person 1, looking at the white board is classified by a yaw rotation angle in zone 11 to 15 and pitch rotation angle in zone 1. Looking at Person 2 is distinguished from Person 3 by a yaw rotation angle in zone 5 to 7 and a pitch rotation angle in zone 2. Hence, the behavior of looking at Person 3 from the point of view of Person 1 is detected by a yaw rotation angle in zone 8 to 10 and a pitch rotation angle in zone 2. In case Person 1 turns back and looks at the screen the viewing zone is unknown because the head pose estimation fails.

Overlapping areas and hence an activity are detected based on the consensus of the participants. If all of the three people look at the white board, we consider the white board as the most informative area for all participants in the meeting room. There is another case: one person gives a presentation in front of the white board and others looks at the him. The close-up cameras cannot capture the presenter so that his viewing zone is unknown. In this case the white board will be the most informative area as well. If two of the three people look at the third one, we consider the third person to speak and therefore to be the most informative area.

TABLE II
ACCURACY OF THE EXTRINSIC PARAMETERS FOR A CLOSE-UP CAMERA

Set	Error in x	Error in y	Error in z	Total Error
1	1.02 cm	20.98 cm	13.66 cm	25.06 cm
2	7.42 cm	17.21 cm	25.55 cm	31.67 cm
3	10.11 cm	19.55 cm	25.56 cm	33.73 cm
4	3.36 cm	9.38 cm	4.30 cm	10.85 cm



(a) (b)

Fig. 5. *Yaw rotation angle and its corresponding zones*. The yaw rotation angle 5a obtained by faceAPI [18] is discretized into eight viewing zones 5b. These zones are used to detect the looking behavior of a particular participant.

C. Detecting the Most Informative View for a Remote Participant

Using the frame-by-frame detections of Section III-B, the question which we are going to answer is: "What is the best view for a remote participant?" This high-level event detection is one particular application of our approach. We hereby assume that views should not switch to frequently for a remote participant. Furthermore, to cope with noisy detections we make a decision for the best view for a certain period T. Within this period we count the frequency of each interesting area and choose the one with the highest frequency.

IV. RESULTS

In order to evaluate our approach, we conducted several experiments and collected over 120 min of video data for meeting analysis. Here, we evaluated every step of Figure 2 separately. In our scenarios three participants were present in all meetings and observed by up to three ambient cameras and three portable close-up cameras (cp. Fig. 1). Each portable close-up camera and two fixed ambient camera, one pointing at a screen and one at a white board, serve as possible video streams for a remote participant.

At first, we evaluate the accuracy of the extrinsic parameters for a close-up camera. To do so, we annotated four different point sets of head positions using the calibrated ambient cameras in a one second interval resulting in $3D$ points for a participant w.r.t. a common coordinate system. Furthermore, we calibrated the close-up camera for comparison. To obtain the extrinsic parameters for the close-up camera, we used the faceAPI [18] to estimate the head pose (position and orientation) w.r.t. the camera coordinate system (Fig. 4). To measure the accuracy of our approach we calculated the Euclidean distance between the calibrated camera center and the estimated camera center of the close-up camera. In Table

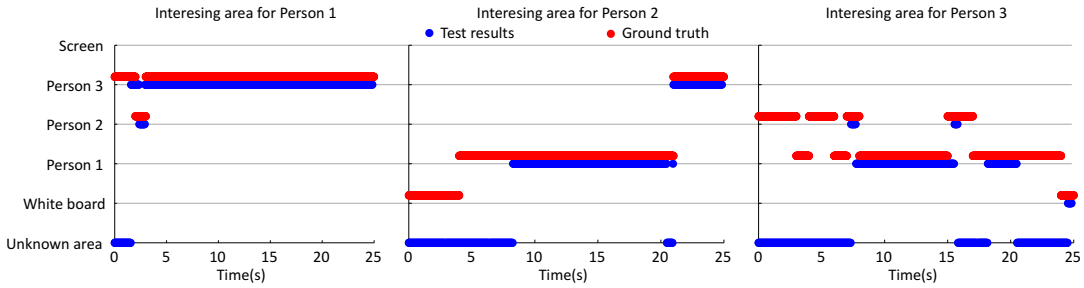


Fig. 6. *Activity detection for each participant*. We annotated a 25 seconds clip for each participant manually in one second intervals. Or approach shows a high accuracy for activity detection of each person except in cases where the activity is unknown. This is due to the fact that faceAPI [18] failed to detect the head pose of this person and therefore a activity cannot be detected.

II we show the results of this calculation in which the accuracy is at most around 35 cm. This is of sufficient accuracy for our particular application.

In the second experiment, we evaluate the detection of interesting areas for several sequences. In Fig. 5a and 5b the results of one complete sequence for Person 2 are shown. The yaw rotation angle is transfered into yaw rotation zones as shown in Fig. 3. To verify the detection for each person we annotated exemplarily a 25 seconds clip of one sequence manually in one second intervals (Fig. 6). In this sequence Person 1 focuses mostly on Person 3. Person 2 looks at the white board at first, then at Person 1 and finally at Person 3. Person 3 on the other hand seems to be active during this sequence due to a alternating focus of Person 1 and Person 3. As a result our algorithm shows high accuracy for activity detection of each person except in cases where the activity is unknown. This is due to the fact that faceAPI [18] failed to detect the head pose of this person and therefore an activity cannot be detected.

Finally, we obtain the overlapping area based on the consensus of all participants (cp. Fig. 6) resulting in an activity. In Fig. 7 the detected activities are compared to manually annotated GT in one second intervals. From the ninth to the sixteenth second, Person 2 and Person 3 look towards Person 1. After that, Person 3 starts talking and becomes the focus. It is possible that people focus on different objects, leading to no overlapping areas. Furthermore, the overlapping area can be unknown due to head pose failure in the faceAPI tracker [18]. Nevertheless, our method is able to find common overlapping areas.

To evaluate the overall performance of our proposed approach, we presented 60 ten-seconds clips out of 120 minutes of recordings to people not involved in this research and asked them to choose a best view out of the presented five video stream. In this context we created ground truth (GT). Anyhow, this manual selection is subjective since it reflects people's decisions and people do not necessarily need to choose the same view. We compared our presented approach to a motion estimation method [22], which measures the overall motion of every view and chooses the one with the highest motion over the ten-seconds period. Although it can be seen in Fig. 8 that the manual selections are quite different, our approach shows

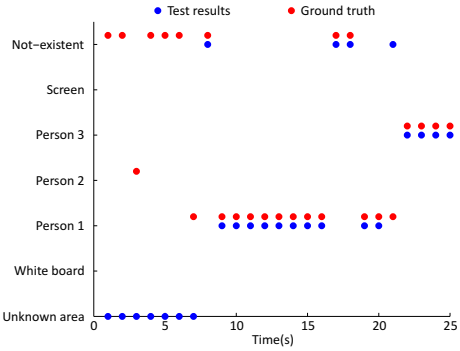


Fig. 7. *Overlapping areas*. We compared the detected activities to manually annotated GT in one second intervals. It is possible that people focus on different objects, leading to no overlapping areas due to head pose failure in the faceAPI tracker [18]. In general, our method shows a robust performance to find common overlapping areas.

clearly a good accuracy. The differences between manual selection and our approach can be explained by the limited features we used to make a decision for the most informative video stream and the looking behavior of people which does not always correspond to the speaker. For the first and second manual selection we achieve a overall performance of 74% in contrast to the motion estimation method which achieves 26%. The third manual selection is quite different from the other GTs which probably can be explained by a different focus of the person who selected the most informative view. Therefore, both approaches achieve a performance of 43%, which is not very accurate.

In summary, our proposed approach performs better than the motion estimation method which measures the overall motion of each view and chooses the one with the highest motion over a ten-seconds period. It is worth to mention that the comparison to manual selections is subjective since it can be biased towards the preference of a certain person. Nevertheless, our approach describes a basic scheme and can be extended by more features to be more robust and detect more activities.

V. Conclusion

In this paper, we presented a novel approach to understand the dynamics in meetings using a multi-camera setup,

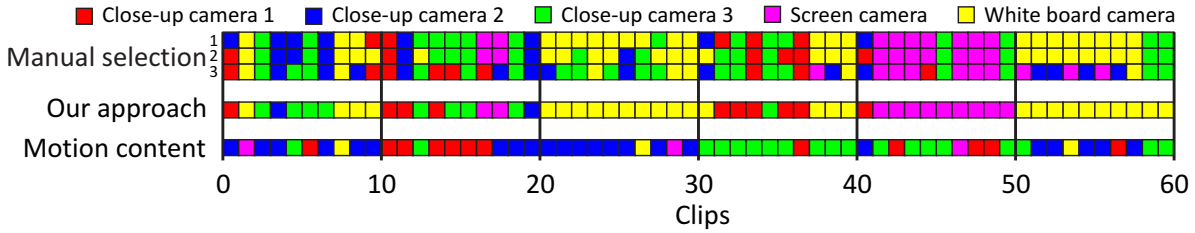


Fig. 8. *Overall performance*. 60 ten-seconds clips out of 120 minutes of recordings were presented to three people not involved in this research for annotation (manual selection 1, 2 and 3). Note that the manual selection is subjective and can be different for every person. Nevertheless, our proposed approach performs better than the motion estimation method (motion content) which measures the overall motion of each view and chooses the one with the highest motion over a ten-seconds period.

consisting of fixed ambient and portable close-up cameras. Our approach is threefold: at first, we estimate the extrinsic parameters of the portable close-up cameras using the head positions. Next, we find common overlapping areas reflecting the looking behavior of people. These areas are found using the head pose (position and orientation) in the close-up cameras. In the third and final step, we detect the most frequent interesting area within a certain time period and use this area to chose the most informative view. We evaluated every step of our proposed approach and showed that our approach performs better than a simple motion estimation method.

There are many possible extensions to this work. One direction could be the incorporation of more features into our approach to increase robustness and to detect more activities. This could potentially lead to more complex applications, for instance creating a complete protocol of a meeting.

ACKNOWLEDGMENT

This research was funded in part by the IBBT iCOCOON and VAU projects co-funded by IBBT (Interdisciplinary institute for Broadband Technology) a research institute founded by the Flemish Government. Companies and organizations involved in the iCOCOON project are Alcatel-Lucent Bell, VITO nv and Eyetronics, with project support of IWT. The work was also sponsored by the Flemish Fund for Scientific Research, through the project "Image and video processing and analysis" (G.0.398.11.N.10).

REFERENCES

[1] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, p. 16, 2011.

[2] S. Grünwedel, V. Jelaca, J. Niño Castañeda, P. Van Hese, D. Van Cauwelaert, P. Veelaert, and W. Philips, "Decentralized tracking of humans using a camera network," in *Proc. of SPIE*, vol. 8301, 2012, p. 9.

[3] S. Gruenwedel, V. Jelaca, P. Van Hese, R. Kleihorst, and W. Philips, "Phd forum: Multi-view occupancy maps using a network of low resolution visual sensors," in *Proc. Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 2011, pp. 1–2.

[4] V. Jelaca, S. Grunwedel, J. Nino-Castaneda, P. Van Hese, D. Van Cauwelaert, P. Veelaert, and W. Philips, "Demo: Real-time indoors people tracking in scalable camera networks," in *Proc. Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 2011, pp. 1–2.

[5] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 267–282, 2008.

[6] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[7] F. Deboeverie, P. Veelaert, and W. Philips, "Face analysis using curve edge maps," in *Lecture Notes in Computer Science*, vol. 6979, 2011, pp. 109–118.

[8] R. Y. Tsai, "Radiometry," in *A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses*, L. B. Wolff, S. A. Shafer, and G. Healey, Eds. Jones and Bartlett Publishers, Inc., 1992, pp. 221–244.

[9] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment a modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. Lecture Notes in Computer Science, B. Triggs, A. Zisserman, and R. Szeliski, Eds. Springer Berlin / Heidelberg, 2000, vol. 1883, pp. 153–177.

[10] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[11] Z. Yu and Y. Nakamura, "Smart meeting systems: A survey of state-of-the-art and open issues," *ACM Computing Surveys*, vol. 42, no. 2, pp. 1–20, 2010.

[12] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L.-w. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. of the tenth ACM International Conference on Multimedia*, ser. MULTIMEDIA '02. ACM, 2002, pp. 503–512.

[13] I. Mikic, K. Huang, and M. Trivedi, "Activity monitoring and summarization for an intelligent meeting room," in *Proc. Workshop on Human Motion*, 2000, pp. 107 –112.

[14] C.-W. Chen and H. Aghajan, "Multiview social behavior analysis in work environments," in *Proc. Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 2011, pp. 1–6.

[15] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing," in *Proc. of the seventh ACM International Conference on Multimedia (Part 1)*, ser. MULTIMEDIA '99. ACM, 1999, pp. 3–10.

[16] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *CHI '02 extended abstracts on Human factors in computing systems*, ser. CHI EA '02. ACM, 2002, pp. 858–859.

[17] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[18] Seeing Machines, "faceapi: http://www.seeingmachines.com/product/faceapi/," May 2009.

[19] K. Arun, T. Huang, and S. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 5, pp. 698–700, 1987.

[20] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[21] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[22] G. Farneback, "Fast and accurate motion estimation using orientation tensors and parametric motion models," in *Proc. of the 15th International Conference on Pattern Recognition*, vol. 1. IEEE, 2000, pp. 135–139.